

CLAIMS

What is claimed is:

1. An apparatus for data compression comprising:

5 an identifier which identifies a plurality of irredundant patterns in a data set; and

an extractor which extracts at least a portion of said plurality of irredundant patterns from
said data set to generate a compressed data set.

2. The apparatus according to claim 1, wherein a more frequently occurring irredundant
10 pattern is extracted before a less frequently occurring irredundant pattern.

3. The apparatus according to claim 1, further comprising:

an ordering device which orders said plurality of irredundant patterns according to a
frequency of occurrence in said data set.

15 4. The apparatus according to claim 1, further comprising:

an input for inputting said data set; and

an output for outputting said compressed data set.

20 5. The apparatus according to claim 1, wherein said at least a portion of said plurality of
irredundant patterns extracted from said data set comprise irredundant patterns having a
minumum frequency of occurrence.

6. The apparatus according to claim 1, wherein an irredundant pattern in said plurality of irredundant patterns comprises a maximal motif, said maximal motif and a location list of occurrences for said maximal motif being incapable of being deduced by a union of a number of location lists of other maximal motifs.

5

7. The apparatus according to claim 6, wherein said maximal motif is maximal in composition and maximal in length.

8. The apparatus according to claim 6, wherein said maximal motif is devoid of a don't care character.

10

9. The apparatus according to claim 1, wherein said data set comprises one of a character string and a character array.

10. The apparatus according to claim 1, wherein said identifier identifies said plurality of irredundant patterns according to an irredundant pattern discovery algorithm.

15

11. The apparatus according to claim 10, wherein said irredundant pattern discovery algorithm comprises:

20

initializing a set of irredundant patterns in said data set;

constructing said set of irredundant patterns for each solid character;

constructing location lists for said set of irredundant patterns, said set of irredundant patterns being iteratively adjusted based on said location lists until no further changes occur to

said set of irredundant patterns; and

updating said set of irredundant patterns.

12. The apparatus according to claim 10, wherein said irredundant pattern discovery
5 algorithm comprises:

computing one-character patterns;

successively growing said one-character patterns by concatenating said one-character
patterns with other patterns;

trimming a number of growing patterns; and

10 using a linearity of 2-motifs to bound a number of said growing patterns.

13. The apparatus according to claim 10, further comprising:

an input for inputting parameters for said irredundant pattern discovery algorithm, said
parameters comprising a string length for said data set, a minimum number of times said
15 irredundant pattern must appear in said data set to be extracted, and a maximum number of
consecutive don't care characters allowed in said irredundant pattern.

14. The apparatus according to claim 1, wherein said data set comprises one of image data,
text data, music data and genetic sequence data.

20 15. The apparatus according to claim 1, wherein said identifier and said extractor comprise a
same device.

16. A facsimile machine comprising the apparatus according to claim 1.

17. A computer comprising the apparatus of claim 1.

5 18. A system for data compression comprising:

an identifying device which identifies a plurality of irredundant patterns in a data set; and

an extracting device which extracts at least a portion of said plurality of irredundant patterns from said data set to generate a compressed data set.

10 19. The system according to claim 18, further comprising:

an input device for inputting said data set;

a memory device for storing said data set; and

an output device for outputting said compressed data set.

15 20. The system according to claim 18, wherein said identifying device identifies said plurality of irredundant patterns according to an irredundant pattern discovery algorithm, said algorithm comprising:

initializing a set of irredundant patterns in said data set;

constructing said set of irredundant patterns for each solid character;

20 constructing location lists for said set of irredundant patterns, said set of irredundant patterns being iteratively adjusted based on said location lists until no further changes occur to said set of irredundant patterns; and

updating said set of irredundant patterns.

21. A data compression/decompression system, comprising:

the data compression apparatus according to claim 1; and

a data decompression apparatus comprising:

an identifier which identifies said irredundant patterns extracted from said data set

in said data compression apparatus; and

an inserter for inserting said extracted irredundant patterns from said data set, into said compressed data set, to reproduce said data set.

22. A method of data compression comprising:

identifying a plurality of irredundant patterns in a data set; and

extracting at least a portion of said plurality of irredundant patterns from said data set to generate a compressed data set.

23. The method according to claim 22, wherein said identifying device identifies said

plurality of irredundant patterns according to an irredundant pattern discovery algorithm, said algorithm comprising:

initializing a set of irredundant patterns in said data set;

constructing said set of irredundant patterns for each solid character;

constructing location lists for said set of irredundant patterns, said set of irredundant patterns being iteratively adjusted based on said location lists until no further changes occur to said set of irredundant patterns; and

updating said set of irredundant patterns.

24. A programmable storage medium tangibly embodying a program of machine-readable instructions executable by a digital processing apparatus to perform a method of data compression, said method comprising:

5 identifying a plurality of irredundant patterns in a data set; and
extracting at least a portion of said plurality of irredundant patterns from said data set to generate a compressed data set.

25. A method for deploying computing infrastructure in which computer-readable code is
10 integrated into a computing system, and combines with said computing system to perform a method of data compression, said method of data compression comprising:

identifying a plurality of irredundant patterns in a data set; and
extracting at least a portion of said plurality of irredundant patterns from said data set to generate a compressed data set.

15